Wide Area Information Servers

Integrated access to personal - corporate - published information

Brewster Kahle
President
WAIS Inc.

What I really want...

- **My personal information to be accessible**
- **Published information should find me**
- **Usable anywhere**
- **Others can use what I have learned
  (if I want them to)**

What is it?

Electronic Publishing

(Or publishing over wires)

New Communications Technology Problems

	BOOKS	Telegraph> Telephone	Electronic Publishing
Experts only	Monks	Operators	Professional searchers
Distribution is hard and expensive	Vellum is calf skin	Telephones on barb wire	$1/minute over obscure modems
Different interfaces	1000's of languages in Europe alone	Switching was manual	//query (W5) inform?
Material is intractable	Scrolls and manuscripts were about as random access as musical scores	No white pages	600 databases on Dialog ~1 Terabyte 140Gbyte at DJ 80GB card catalog at RLG
Business model needed	Centralized printing	Pay per minute	Not understood

Navigation Techniques: Paper

- Alphabetical Listings (dictionary, Encyclopedia)

- Indices (back of the book and Readers Guide)

- Table of Contents (outlining)

- Citation index

- "Tree of Knowledge"

- Have you read any good books lately?

Navigation Techniques: Computers

- Hierarchical File Systems

- Unix "find" and "grep", Mac "find file"

- Gopher, Magellan, ON Location

- Boolean query systems (...within 5 words of...)

- Static Hypertext links (see also pointers)

Navigation Techniques: WAIS

- **English language questions and relevance feedback**
  - **Question-answer dialog**
  - **Similar to Newspapers: "More on page 5"**
  - **Dynamic Hypertext Links**

- **2 level search:**
  - **Directory of servers (server like any other)** -
  Servers themselves

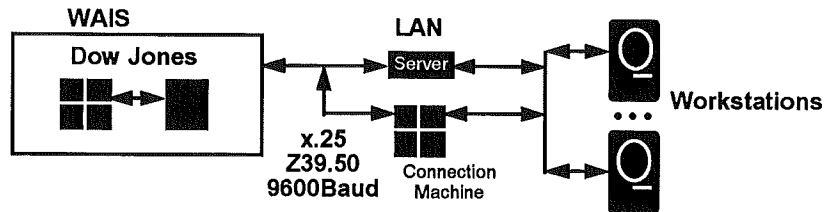---

Wide Area Information Server Architecture
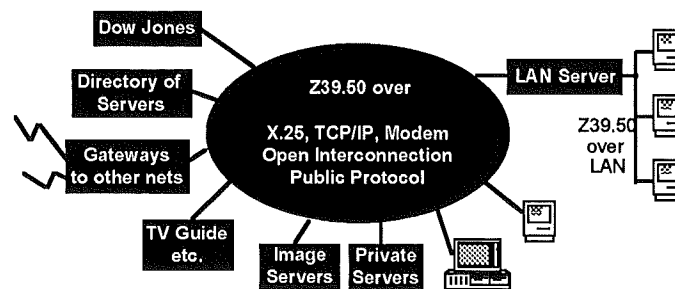
Users Needs:
Selecting Servers
Answering Questions
Organizing Responses

Architecture Issues:
Scalability
Security
Business model for servers
Reliable Access

Demonstration System Structure

WAIS

Dow Jones

LAN

Server

x.25
Z39.50
9600Baud   Connection
Machine

Workstations

Operations:
 Archiving
 Queries
 Retrieval
IR Type:
 Broadcast
 Query by Example
Databases:
 Wall St Journal
 Barron's
 400 Business Mags

CM:
Operations:
  Queries
IR Type:
  enhanced relevance
  feedback
Databases:
  DowVision and memo's,
  mail, word processor files

Mac:
Operations:
  Human Int
  Retrieval
  Queries
  "Caching" Docs
  User Profiles
IR Type:
  Query by example
Databases:
  Personal Text
  Cached data

WAIS Clients

* Busy 24 hours a day finding information

* Ponder all indications of the preferences of its user

* Gossip with other clients about their discoveries

* Scours the world (within a budget) to find
  new sources

* Current implementations on PC, Macintosh,
  X Windows, NeXT, dumb terminal (dial-up)

WAIS Protocol

- Based on NISO Z39.50 international standard
- Flexible — separates clients from servers
- Search: *(words, doc_ids, databases)*
  returns list of: *(headline, score, doc_id, types)*
- Retrieval: *(doc_id, type, start, end)* returns:
  data of specified type
- Doc_id: An ISBN for the Electronic Age
- Server Description Structure for the Directory
  of Servers

How Standard Protocol
can Provides Security

- Users do not login to server, but search only
  through application layer protocol (Z39.50).
- Server controls access to data.
- Network layers below application, or application
  layer handles authentication, encryption, billing.

The WAIS Protocol *is* WAIS

Client — Z39.50 — Server
Server
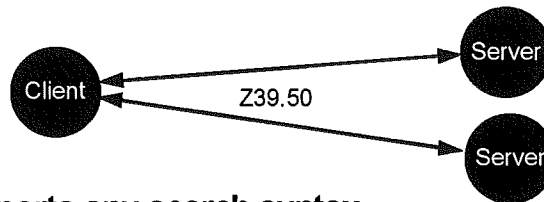
- Supports any search syntax
- Supports sophisticated clients —   puts
  intelligence in the user's hands
- Clients can run on any platform
- Multiple servers in a single search
- Retrieve any kind of data: text, graphics, video,...

Connection Machine Server

- 1-100GBytes (and getting bigger)
- Supports thousands of users
- Automatic Indexing
- Uses words and phrases in question to find
  appropriate documents with relevance feedback,
  weighted term
- Supports Boolean Queries
- Cost effective hardware alternative to mainframes

Unix Based Server

- 1MB-10GBytes
- Supports hundreds of users
- Multimedia indexing: word processor files etc.
- Supports Natural language and Boolean Queries
- Easy to install

Wide Area Information Servers

Results Improve with Query Size

Precision x recall
@ 25% recall

Average
performance
over 13
reference sets

Typical
Boolean
Query

Typical
Relevance
Feedback
Query

.6
.5
.4
.3
.2
.1
0

10 20 30 40 50 60 70 80 90 100

number of query terms

Wide Area Information Servers
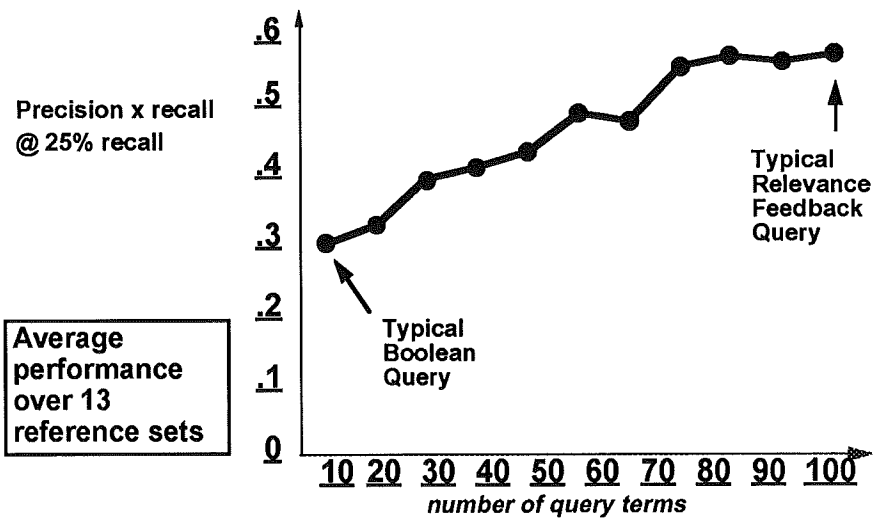
WAIStation:
Active Database Sources,
Saved Questions

Sources		Questions
◉ CM applications		? CM Apps Question
◉ Encyclopedia		? Library question
◉ King James Bible		? Encyclopedia Q
◉ Macintosh Hard Disk		? Patent Q
◉ TMC Business email		? TMC Bus. Email Q
◉ TMC Library Catalog		? TMC Fun Q
◉ *Wall St. Journal*		? Montvale Q
◉ World Factbook		? World Factbook Q
		? poetry q
		? Bible Q

Select Data Source

Sources — **Question-1**

Look for documents about

[Run]

Which are similar to in these sources
◉ *Wall St. Journal*
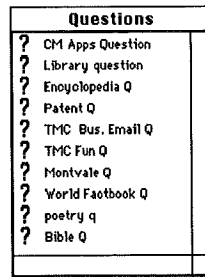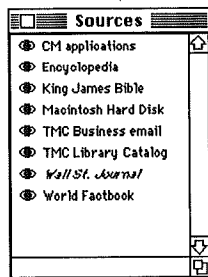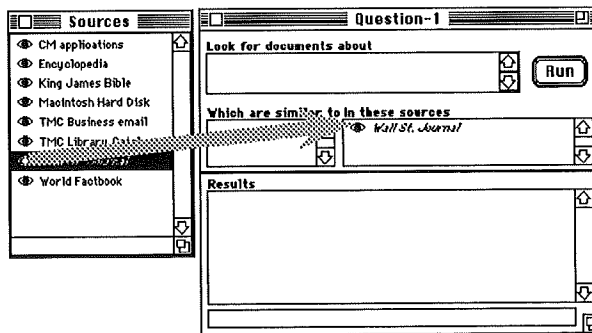
Results

Run Initial Query

```
▉▉▉▉▉▉▉▉▉▉ Question-1 ▉▉▉▉▉▉▉▉▉▉

Look for documents about
┌─────────────────────────────────┐ ┌─┐ ┌───────┐
│ recent developments in personal │ │△│ │  Run  │
│ computers│                       │ │▽│ └───────┘
└─────────────────────────────────┘ └─┘

Which are similar to in these sources
┌──────────────┐ ┌─┐ ┌ ● Wall St. Journal ──────┐ ┌─┐
│              │ │△│ │                           │ │△│
│              │ │▽│ │                           │ │▽│
└──────────────┘ └─┘ └───────────────────────────┘ └─┘

Results
┌──────────────────────────────────────────────────┐┌─┐
│ 🖹 *** Compaq Computer Directors Approve 2-for-1 Stock Split ││△│
│ 🖹 *** International: Bull Agrees to Pay Zenith $15 Million to End ││▓│
│ 🖹 *** AT&T Set to Announce Memorex Computer Accord ││▓│
│ 🖹 *** Technology Brief -- International Business Machines: Pri ││▓│
│ 🖹 *** Business Brief -- Data General Corp.: Four Models Are Un ││▓│
│ 🖹 *** Technology: Computer Firms See the Writing on the Screen ││▓│
│ 🖹 *** Retailing: Businessland Enters Japan, Aided by 4 Big Loca ││▓│
│ 🖹 *** Corrections & Amplifications ││▽│
└──────────────────────────────────────────────────┘
┌──────────────────────────────────────────────────┐ ▣
└──────────────────────────────────────────────────┘
```

Click a Headline to Display a Document

```
▉▉▉▉▉ Question-1 ▉▉▉▉▉

Look for documents about
┌────────────────────────┐ ┌─┐ ┌──────┐
│ recent developments in personal │ │△│ │ Run │
│ computers│              │ │▽│ └──────┘
└────────────────────────┘ └─┘

Which are similar to in these sources
┌────────┐ ┌─┐ ┌ ● Wall St. Journal ──┐ ┌─┐
│        │ │△│ │                       │ │△│
│        │ │▽│ │                       │ │▽│
└────────┘ └─┘ └───────────────────────┘ └─┘

Results
┌──────────────────────────────────────┐┌─┐
│ 🖹 *** Compaq Computer Directors Approve 2-for-1 Stock Split ││△│
│ 🖹 *** International: Bull Agrees to Pay Zenith $15 Million to En ││▓│
│ 🖹 *** AT&T Set to Announce Memorex Computer Accord ││▓│
│ 🖹 *** Technology Brief -- International Business Machines: Pri ││▓│
│ 🖹 *** Business Brief -- Data General Corp.: Four Models Are Un ││▓│
│ 🖹 *** ▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓ ││▓│
│ 🖹 *** Ret┌─▉ Technology: Computer Firms See the Writing ▉──┐
│ 🖹 *** Cor│  International Business Machines Corp., Apple Computer Inc. │△│
└───────────│  and other big computer makers are staking out positions in │▓│
            │  the nascent market for "note-pad computers," small machines │▓│
            │  that let users enter data by writing rather than typing     │▓│
            │  keys. The note pads typically recognize numbers and letters │▓│
            │  printed on a screen with a special pen and convert them into│▓│
            │  conventional electronic characters. The information is then │▓│
            │  stored for later transfer to a personal computer or a       │▓│
            │  company's main computers.                                   │▓│
            │     The size of the market for note-pad computers isn't clear,│▓│
            │  but Infocorp, a Santa Clara, Calif., market-research firm,  │▓│
            │  estimates the market will grow to 5.4 million units sold in │▓│
            │  1995 from 22,000 units this year. Only one company, Tandy   │▓│
            │  Corp.'s Grid Systems unit, currently sells note-pad computers│▓│
            │  in the U.S.; its model, introduced last September, is priced│▓│
            │  at $3,000. But new ventures are expected to introduce several│▓│
            │  note-pad machines this year. And already, big computer makers│▓│
            │  are fighting quietly for control over software standards for │▓│
            │  these gadgets, which require different programs from those   │▽│
            └──────────────────────────────────────────────────────────────┘
```

Relevance feedback:
"Find me more like this one"

Question-1

Look for documents about

recent developments in personal
computers
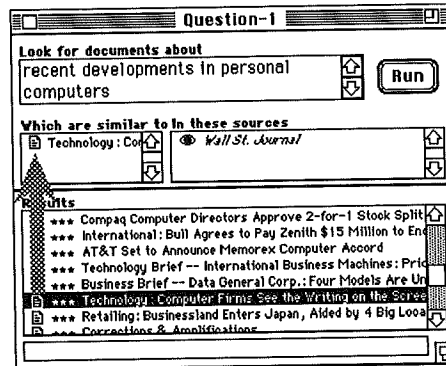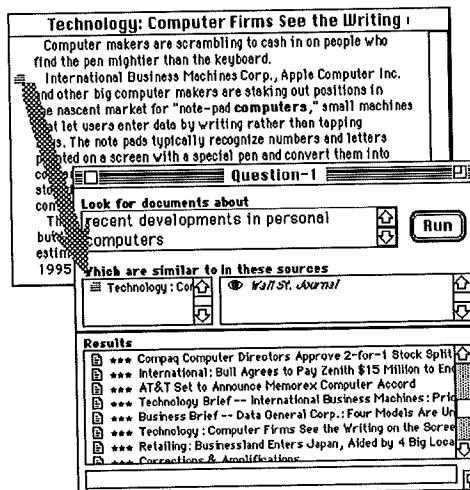
(Run)

Which are similar to in these sources

🖹 Technology : Co   ● *Wall St. Journal*

Results

*** Compaq Computer Directors Approve 2-for-1 Stock Split
*** International: Bull Agrees to Pay Zenith $15 Million to En
*** AT&T Set to Announce Memorex Computer Accord
*** Technology Brief -- International Business Machines: Pri
*** Business Brief -- Data General Corp.: Four Models Are U
*** Technology: Computer Firms See the Writing on the Scree
*** Retailing: Businessland Enters Japan, Aided by 4 Big Loca
*** Corrections & Amplifications

Relevance Feedback of Paragraph

Technology: Computer Firms See the Writing

Computer makers are scrambling to cash in on people who
find the pen mightier than the keyboard.
International Business Machines Corp., Apple Computer Inc.
nd other big computer makers are staking out positions in
e nascent market for "note-pad **computers**," small machines
t let users enter data by writing rather than tapping
s. The note pads typically recognize numbers and letters
ted on a screen with a special pen and convert them into

Question-1

Look for documents about

recent developments in personal
omputers

(Run)

Which are similar to in these sources

≡ Technology : Co   ● *Wall St. Journal*

Results

*** Compaq Computer Directors Approve 2-for-1 Stock Split
*** International: Bull Agrees to Pay Zenith $15 Million to En
*** AT&T Set to Announce Memorex Computer Accord
*** Technology Brief -- International Business Machines: Pri
*** Business Brief -- Data General Corp.: Four Models Are Ur
*** Technology : Computer Firms See the Writing on the Scree
*** Retailing: Businessland Enters Japan, Aided by 4 Big Loca
*** Corrections & Amplifications

"Chaining" of Questions to Follow a Tangent

Question-1

Look for documents about

recent developments in personal computers

[Run]

Which are similar to in these sources

≡ Technology : Co   👁 *Wall St. Journal*

Results
- ••• International: Bull Agrees to Pay Zenith $15 Million to En
- ••• AT&T Set to Announce Memorex Computer Accord
- ••• Technology Brief -- International Business Machines: Pri
- ••• Business Brief -- Data General Corp.: Four Models Are Un
- ••• Technology : Computer Firms See the Writing on the Scre

Question-2

Look for documents about

[Run]

Which are similar to in these sources

Retailing: Busin   👁 *Wall St. Journal*

Results
- +++ Retailing: Businessland Enters Japan, Aided by 4 Big Loca
- ++ What's News -- Business and Finance
- ++ Technology: Computer Makers Agree on a Standard For N
- + Inside Track: Businessland Directors Take a Loss And Tra
- + Technology & Health: Businessland To Report Loss For 3r
- + Technology: U.S. Computer Maker Takes on NEC on Its Ov
- + Technology: Computer Firms See the Writing on the Scre

Internet Freeware Release

- **Example User interfaces for free (no support) Macintosh, Gnu Emacs, Xwindows**

- **Example unix server software to create servers**

- **Directory of Servers on the internet at least through '92**

- **300 Servers now: Weather Maps, patents, journal abstracts, email archives, usenet recipies,...**

- **Free Software via FTP from Think.com:/wais/***

  Mailing list: wais-discussion-request@think.com

WAIS Uses

- Over 15,000 users on the Internet
- Users in 28 Countries: Mexico, Singapore, Finland, Australia, etc
- 300 Databases served from 12 Countries: Norway, Canada, UK, etc. Average 3 new databases registered per week.

WAIS Uses:
Campus Wide Info Servers

- Class catalog and schedule
- Campus events: movies, sports
- Job listings
- Library catalog
- Phone book
- Professor research interests
- Past theses

```
[     sol.acs.unt.edu]   UNTComputerDoc
[        xantos.uio.no]   UiO_Publications
[    next2.oit.unc.edu]   ibm.pc.FAQ
```

WAIS Uses: Libraries

- **Easy to use card catalog**
- **Remote use from home or office**
- **Pictures, full text, scanned documents**

```
[pegun.law.columbia.e]  columbia-law-library-catalog
[pegun.law.columbia.e]  columbia-spanish-law-catalog
[      quake.think.com]  tmc-library
```

WAIS Uses: Biology

- **Journal Abstracts**
- **Sequence archives**
- **Images**

Currently over 20 Biology databases in Finland, Netherlands, and US

```
[      cmns.think.com]  Molecular-biology
[          bio.vu.nl]  biology-compounds
[     genbank.bio.net]  biology-journal-contents
[       wais.funet.fi]  bionic-ai-researchers
[       wais.funet.fi]  bionic-directory-of-servers
[       wais.funet.fi]  bionic-enzyme
```

WAIS Uses: Chemistry
CORE Project

- **All published chemistry (8 years all ACS)**
- **Scanned pictures, ascii text**
- **Optical jukebox mass storage**
- **Connection Machine / Newton search engines**

**Project of Bellcore, ACS, Chem Abstracts,
OCLC, Cornell, and Thinking Machines**

`[  cujo.curtin.edu.au]  chem-eng-current-contents`

WAIS Uses:
Business Executives

- **Dow Jones information**
- **Corporate information**
- **Personal information**

**Project:
KPMG, Apple, Thinking Machines, Dow Jones**

```
[       cmns.think.com]  wall-street-journal-sample
[            think.com]  Business-email
```

WAIS Uses:
Medical Researchers/Doctors

- **Medical papers**
- **Storing and matching patient records**
- **Remote connections to specialized databases**

```
[       wais.funet.fi]  bionic-databases-limb
```

WAIS Uses:
Community Information

- **Dial-up users: no network required**
- **Directories of services or facilities**
- **Education and entertainment**

```
[     quake.think.com]  internet-resource-guide
[     sol.acs.unt.edu]  online-libraries
[     quake.think.com]  weather
[ lambada.oit.unc.edu]  nsf-bulletins
```

Conclusion

- Electronic Publishing can fill niches now.
- WAIS Inc is working with groups that want to publish their information internally or externally.
- WAIS Inc is working with those that want to profit by using other peoples servers.

WAIS Hardware Components
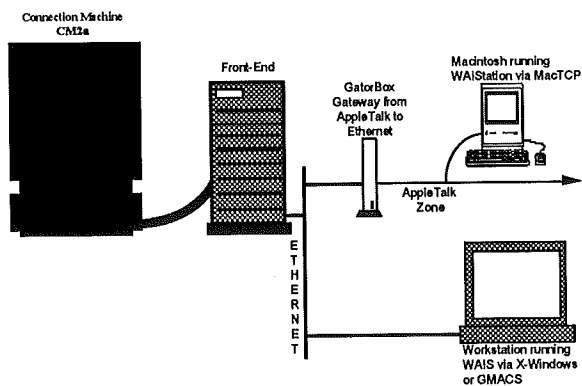
Data Parallelism:
Searching all the Documents at Once

Pharmaceutical +12

FDA +9

Medical +6
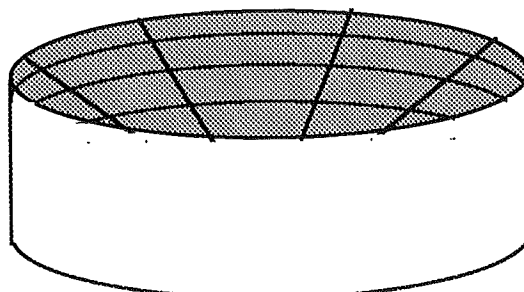
Stadium

Boolean Search   Conceptual Search
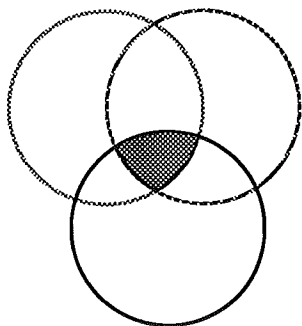
Retrieve documents containing
specific combinations of words
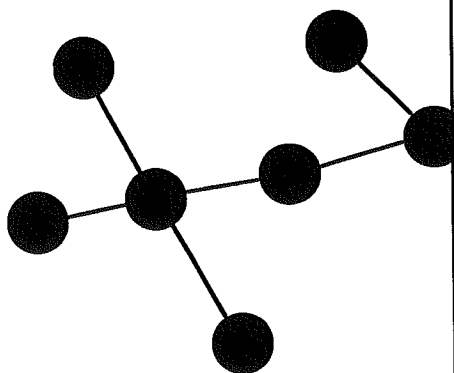
Explore a set of documents
containing related concepts

Boolean Query

Hard to Use:
Complex Syntax

(Japanese OR Japan) AND
(building OR buildings OR (Real AND Estate) AND
(Manhattan OR (New AND York)

Poor Results:
The wrong information
No ranking of results

Have you been paying attention?...
Freer Finance: U.S. Regulators Move...
REAL ESTATE: California Initiatives...
First Boston Said To Agree on Sale Of...
Exxon, Rockefeller Group to Sell Site...
What's News--Business and Finance

Conceptual Search: Phase 1

Easy to Use:
No Syntax

Japanese buying real estate in mid-town manhattan

Options:
What do you want
to follow up?

1. Time Acts to Cut Magazine Costs...
2. First Boston Said To Agree on Sale...
3. Have You Been Paying Attention?
4. Exxon, Rockefeller Group to Sell Site...
5. Hard Sell: Real Estate Developers...
6. What's News--Business and Finance...
7. Integrated Resources Buys Loft Building...

Conceptual Search: Phase 2

**Relevance Feedback:**
I like these;   show me more.

First Boston Said To Agree on Sale...

Exxon, Rockefeller Group to Sell Site...

Improved results:

Articles on related topics are found.

Results are ranked.

1. Bids for Exxon Building in New York...
2. Time Acts to Cut Magazine Costs...
3. Hard Sell: Real Estate Developers...
4. Time Inc. Sells Its 45% Interest...
5. Citicorp Unit Moves to Foreclose on...
6. Litigious Landlords: Legal Maneuvers

Query Broadcast To Database on Connection Machine System

Document Units   Scores

Tripoli

Libyan

PLO

bomb

75

42

80

52

17
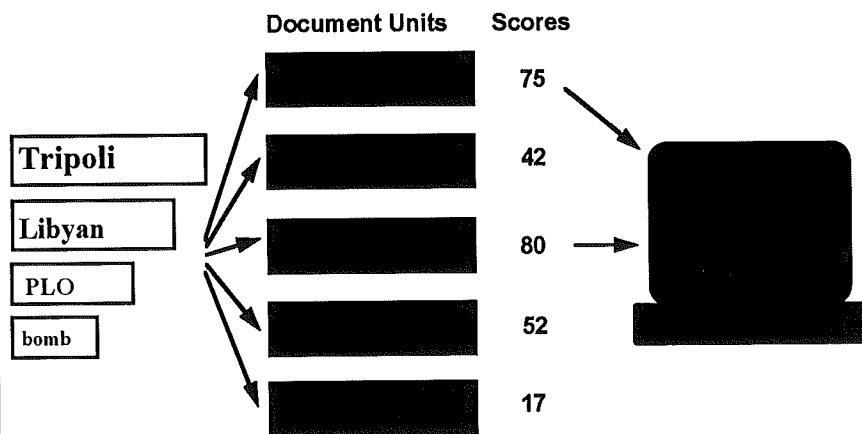
Document Retrieval Performance

- **Current algorithm limits:**
  ~2 GB with 512 MB CM-2
  ~8 GB with 2 GB CM-2
  ~25 GB with 8 GB CM-2

- **High recall**   } see   Stanfill and Kahle,
                                        Communications of the ACM,
- **High precision**   December 1986

- **<< 1 sec. response**

- **Much larger DBs searchable with CM-5**
  and inverted index algorithms: 100s to 1000s of
  Gigabytes